A Simple Modified Verification Method for Linear Systems

Atsushi Minamihata(Waseda Univ.)
joint work with
Kouta Sekine,Takeshi Ogita,
Siegfried M. Rump, Shin'ichi Oishi

aminamihata@moegi.waseda.jp

SCAN 2014

September 24, 2014

Object Problem

For linear system,

$$Ax = b, \tag{1}$$

we can efficiently obtain a computed solution $\tilde{x}$ by some numerical algorithm. In general, however, we do not know how accurate the computed solution is.

Aim

In this talk, we aim to calculate error bounds of $\tilde{x}$ of $Ax = b$ to the exact solution $x^* := A^{-1}b$ such that

$$|\tilde{x}_i - x_i^*| \leq \epsilon_i, \quad (i = 1, 2, \cdots, n)$$

by the use of verified numerical computations.

previous work

Several Verification Method

Several Verification Method have been proposed. For example,

- Neumaier [1]
- Rump [2]

In particular, we'll consider [2].

[1] : A. Neumaier. A simple derivation of the Hansen-Bliek-Rohn-Ning-Kearfott enclosure for linear interval equations. Reliable Computing, 5:131—136, 1999
[2] : S. M. Rump, Accurate solution of dense linear systems, Part II: Algorithms using directed rounding, J. Comp. Appl. Math., 242 (2013), 185–212.

Propose

- Modification of Rump's method [2]
  - We propose two simple modified methods.

Rump's method

- Implementation as verifylss in INTLAB 7
- Computational complexity is $\mathcal{O}(n^2)$

INTLAB

The Matlab toolbox for reliable computing and self-validating algorithms.

Notation

- Let I denote the $n \times n$ identity matrix
- Let O denote the $n \times n$ matrix of all zeros

For real $n \times n$ matrices $A = (a_{ij})$ and $B = (b_{ij})$,

- The notation $A \leq B$ means $a_{ij} \leq b_{ij}$ for all (i, j)
- The notation $|A|$ means $|A| = (|a_{ij}|) \in \mathbb{R}^{n \times n}$

Similar notation is applied to real vectors.

Definition

monotone

A real $n \times n$ matrix A is called monotone if $Ax \geq \mathbf{0}$ implies $x \geq \mathbf{0}$ for $x \in \mathbb{R}^n$.

Z-matrix

Let $A = (a_{ij})$ be a real $n \times n$ matrix with $a_{ij} \leq 0$ for $i \neq j$. Then A is called a Z-matrix.

M-matrix

If a Z-matrix A is monotone, then A is called an M-matrix.

Definition

comparison matrix

The comparison matrix $\langle A \rangle = (\widehat{a}_{ij})$ of A is defined by

$$\widehat{a}_{ij} = \begin{cases} |a_{ij}| & (i = j) \\ -|a_{ij}| & (i \neq j) \end{cases} .$$

H-matrix

If $\langle A \rangle$ is an M-matrix, then A is called an H-matrix.

Property of M-matrix and H-matrix

Theorem (Fiedler-Pták)

Let an Z-matrix $A \in \mathbb{R}^{n \times n}$ be given. Then the following conditions are equivalent:

1. *A is nonsingular and $A^{-1} \geq O$ (i.e., A is an M-matrix).*
2. *There exists $v \in \mathbb{R}^n$ with $v > \mathbf{0}$ satisfying $Av > \mathbf{0}$.*

Theorem (well-known)

If A is an H-matrix, then

$$|A^{-1}| \leq \langle A \rangle^{-1}.$$

Theorem 1 (Rump [2, Theorem 2.1])

Let $A \in \mathbb{R}^{n\times n}$ and $b, \tilde{x} \in \mathbb{R}^n$ be given. Assume $v \in \mathbb{R}^n$ with $v > \mathbf{0}$ satisfies $u := \langle A\rangle v > \mathbf{0}$. Let $\langle A\rangle = D - E$ denote the splitting of $\langle A\rangle$ into the diagonal part D and the off-diagonal part $-E$, and define $w \in \mathbb{R}^n$ by

$$w_k := \max_{1\le i\le n} \frac{G_{ik}}{u_i} \quad \text{for } 1 \le k \le n,$$

where $G := I - \langle A\rangle D^{-1} = ED^{-1} \ge O$. Then A is nonsingular, and

$$|A^{-1}b - \tilde{x}| \le (D^{-1} + vw^T)|b - A\tilde{x}|. \tag{2}$$

Proof

From the definition of u and w, it follows

$$O \leq I - \langle A \rangle D^{-1} \leq uw^T. \tag{3}$$

Multiplying (3) from the left by $\langle A \rangle^{-1}$ yields

$$\langle A \rangle^{-1} - D^{-1} \leq \langle A \rangle^{-1} uw^T = \langle A \rangle^{-1} \langle A \rangle vw^T$$

and

$$\langle A \rangle^{-1} \leq D^{-1} + vw^T. \tag{4}$$

Using $|A^{-1}| \leq \langle A \rangle^{-1}$ and (4),

$$\begin{aligned} |A^{-1}b - \tilde{x}| &\leq |A^{-1}||b - A\tilde{x}| \leq \langle A \rangle^{-1} |b - A\tilde{x}| \\ &\leq (D^{-1} + vw^T)|b - A\tilde{x}|. \end{aligned}$$

Modification of Rump's theorem

Theorem 2

Let $A, b, \tilde{x}, u, v, w$ be defined as in Theorem 1. Define $D_s := \mathrm{diag}(s)$ where $s \in \mathbb{R}^n$ with

$$s_k := u_k w_k \quad \text{for } 1 \leq k \leq n.$$

Then,

$$|A^{-1}b - \tilde{x}| \leq (D^{-1} + vw^T)(I + D_s)^{-1}|b - A\tilde{x}|. \tag{5}$$

Moreover,

$$|A^{-1}b - \tilde{x}| \leq \beta v + (D^{-1} + vw^T)(I + D_s)^{-1}(|b - A\tilde{x}| - \beta u), \tag{6}$$

where $\beta := \min_{1 \leq i \leq n} \frac{|b - A\tilde{x}|_i}{u_i}$.

Proof

From the definition of u and w, it follows

$$I - \langle A \rangle D^{-1} \leq uw^T.$$

Since $\mathrm{diag}(I - \langle A \rangle D^{-1}) = \mathbf{0}$, we have

$$I - \langle A \rangle D^{-1} + D_s \leq uw^T$$

and

$$I + D_s \leq \langle A \rangle D^{-1} + uw^T. \tag{7}$$

Proof

From the assumption, A is an H-matrix, so that $\langle A \rangle^{-1} \geq O$.
Multiplying (7) from the left by $\langle A \rangle^{-1}$ yields

$$\langle A \rangle^{-1}(I + D_s) \leq D^{-1} + \langle A \rangle^{-1} uw^T = D^{-1} + vw^T.$$

Since $(I + D_s)^{-1} \geq O$, we have

$$\langle A \rangle^{-1} \leq (D^{-1} + vw^T)(I + D_s)^{-1}. \tag{8}$$

Using $|A^{-1}| \leq \langle A \rangle^{-1}$ and (8),

$$\begin{aligned} |A^{-1}b - \tilde{x}| &\leq |A^{-1}||b - A\tilde{x}| \leq \langle A \rangle^{-1}|b - A\tilde{x}| \\ &\leq (D^{-1} + vw^T)(I + D_s)^{-1}|b - A\tilde{x}|. \end{aligned}$$

Proof

From the definition of β and the property of H-matrix,

$$\begin{aligned} |A^{-1}b - \tilde{x}| &\leq |A^{-1}||b - A\tilde{x}| \\ &\leq \langle A \rangle^{-1}|b - A\tilde{x}| \\ &= \beta\langle A \rangle^{-1}u + \langle A \rangle^{-1}(|b - A\tilde{x}| - \beta u) \\ &= \beta v + \langle A \rangle^{-1}(|b - A\tilde{x}| - \beta u). \end{aligned}$$

Using (8),

$$\begin{aligned} |A^{-1}b - \tilde{x}| &\leq \beta v + \langle A \rangle^{-1}(|b - A\tilde{x}| - \beta u) \\ &\leq \beta v + (D^{-1} + vw^T)(I + D_s)^{-1}(|b - A\tilde{x}| - \beta u). \end{aligned}$$

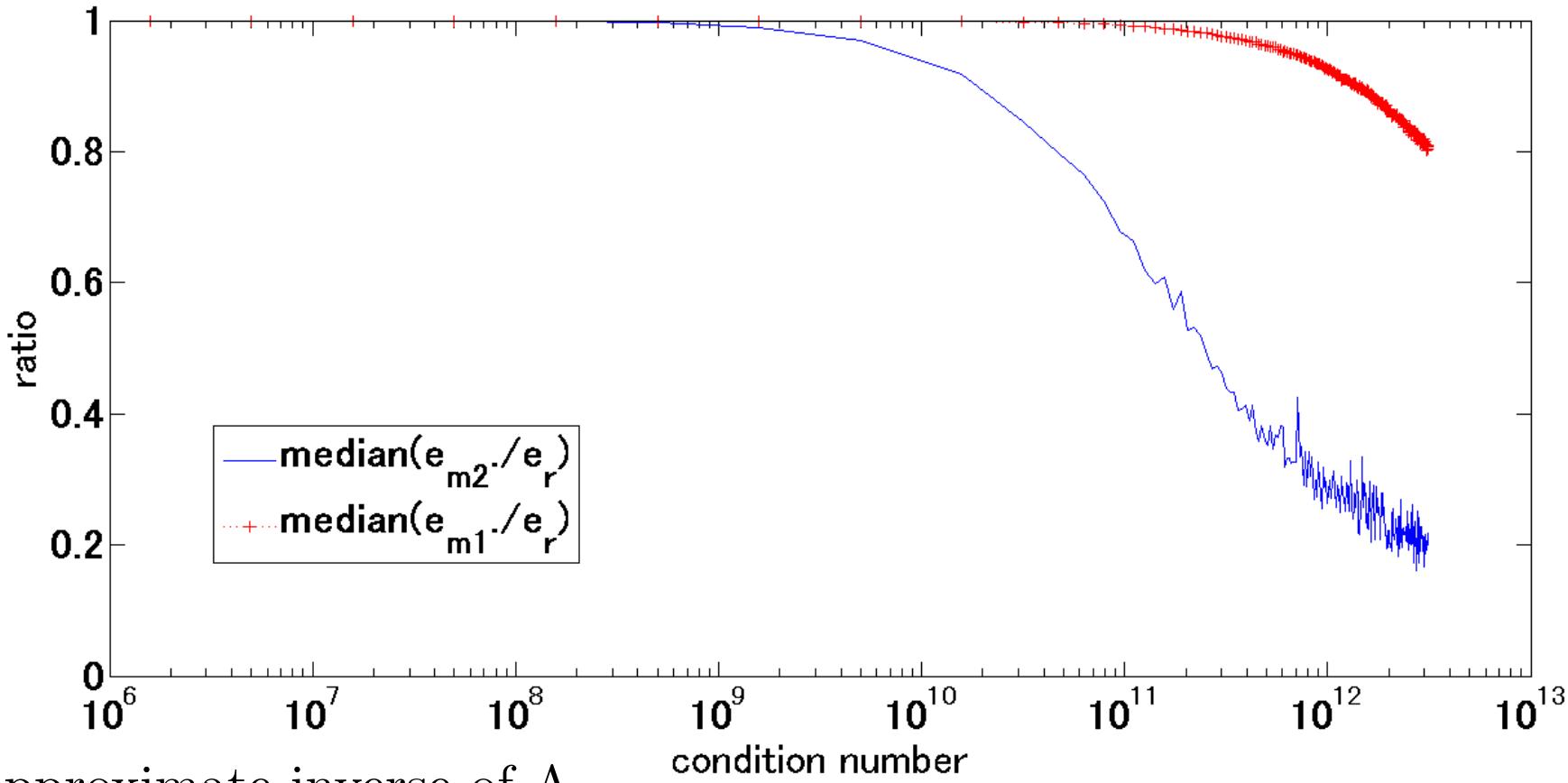
Numerical Result

- Test Matrix : randmat function (INTLAB function)
- Matrix Size : 100

- ϵ_r : Error bounds based on (2) in Theorem 1
  - $|A^{-1}b - \tilde{x}| \leq (D^{-1} + vw^T)|b - A\tilde{x}|$
- ϵ_{m1} : Error bounds based on (5) in Theorem 2
  - $|A^{-1}b - \tilde{x}| \leq (D^{-1} + vw^T)(I + D_s)^{-1}|b - A\tilde{x}|$
- ϵ_{m2} : Error bounds based on (6) in Theorem 2
  - $|A^{-1}b - \tilde{x}| \leq \beta v + (D^{-1} + vw^T)(I + D_s)^{-1}(|b - A\tilde{x}| - \beta u)$

Numerical Result

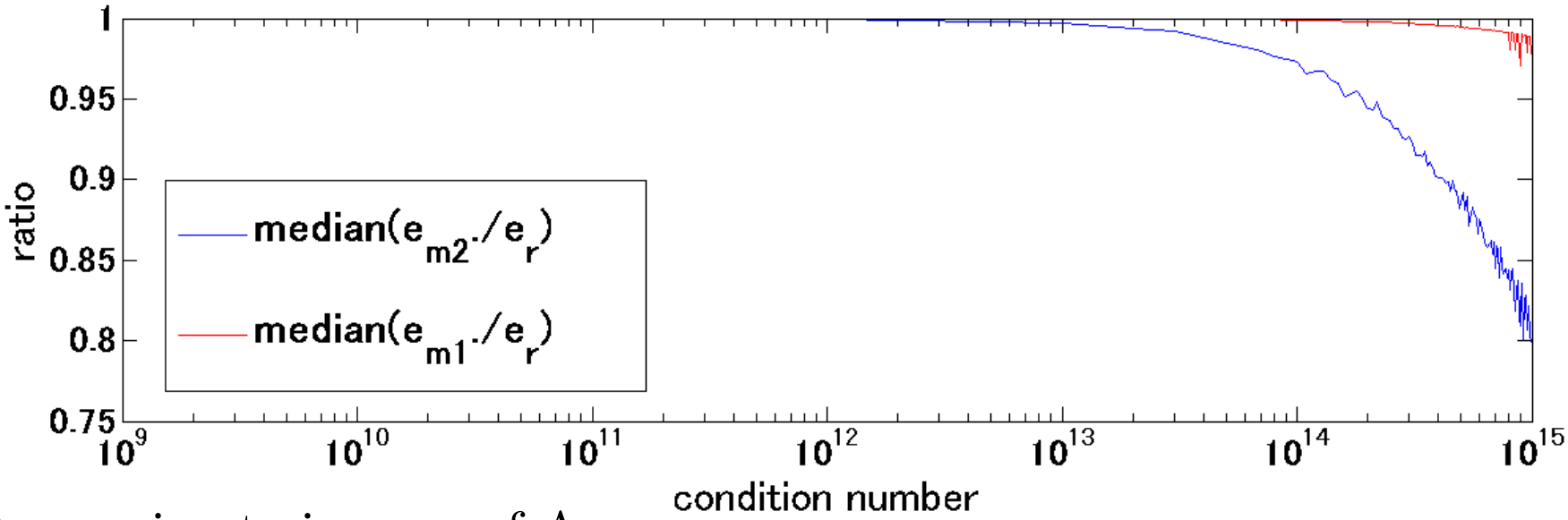
- The inclusion of RA :
  $[\mathrm{fl}_\nabla(RA), \mathrm{fl}_\nabla(RA) + \gamma_{2n}|R||A| + n * 2^{-1022} * ee^T]$
- ratio := median$(\epsilon_M./\epsilon_R)$



R : Approximate inverse of A
$\mathrm{fl}_\nabla(RA)$: setround$(-1); R * A$

Numerical Result 2

- The inclusion of RA : $[\mathrm{fl}_{\triangledown}(RA), \mathrm{fl}_{\triangle}(RA)]$
- ratio := median$(\epsilon_M./\epsilon_R)$



R : Approximate inverse of A

$\mathrm{fl}_{\triangledown}(RA)$: setround$(-1); R * A$, $\mathrm{fl}_{\triangle}(RA)$: setround$(1); R * A$

Conclusion 1

Theorem 2 requires $\mathcal{O}(n)$ floating-point operations more than original theorem.

- The inclusion of RA :
  $[\mathrm{fl}_{\triangledown}(RA), \mathrm{fl}_{\triangledown}(RA) + \gamma_{2n}|R||A| + n * 2^{-1022} * ee^T]$
  - We improved accuracy.
- The inclusion of RA : $[\mathrm{fl}_{\triangledown}(RA), \mathrm{fl}_{\triangle}(RA)]$
  - We improved accuracy a little.

Theorem 3

Let $A, b, \tilde{x}, u, v, w$ be defined as in Theorem 1. Define $\Delta := uw^T - ED^{-1}$ and $c := |b - A\tilde{x}|$. Then,

$$|A^{-1}b - \tilde{x}| \leq (D^{-1} + vw^T)(c - \Delta(I - \Delta)c).$$

Moreover,

$$|A^{-1}b - \tilde{x}| \leq \beta v + (D^{-1} + vw^T)((c - \beta u) - \Delta(I - \Delta)(c - \beta u)),$$

where $\beta := \min_{1 \leq i \leq n} \frac{c_i}{u_i}$.

Theorem 3

Let $A, b, \tilde{x}, u, v, w$ be defined as in Theorem 1. Define $\Delta := uw^T - ED^{-1}$ and $c := |b - A\tilde{x}|$. Then,

$$|A^{-1}b - \tilde{x}| \leq (D^{-1} + vw^T)(c - \Delta(I - \Delta)c).$$

Moreover,

$$|A^{-1}b - \tilde{x}| \leq \beta v + (D^{-1} + vw^T)((c - \beta u) - \Delta(I - \Delta)(c - \beta u)),$$

where $\beta := \min_{1 \leq i \leq n} \frac{c_i}{u_i}$.

$$|A^{-1}b - \tilde{x}| \leq \beta v + (D^{-1} + vw^T) \min((c - \beta u), ((c - \beta u) - \Delta(I - \Delta)(c - \beta u)))$$

Theorem 3

Let $A, b, \tilde{x}, u, v, w$ be defined as in Theorem 1. Define $\Delta := uw^T - ED^{-1}$ and $c := |b - A\tilde{x}|$. Then,

$$|A^{-1}b - \tilde{x}| \leq (D^{-1} + vw^T)(c - \Delta(I - \Delta)c). \tag{9}$$

Moreover,

$$|A^{-1}b - \tilde{x}| \leq \beta v + (D^{-1} + vw^T)((c - \beta u) - \Delta(I - \Delta)(c - \beta u)), \tag{10}$$

where $\beta := \min_{1 \leq i \leq n} \frac{c_i}{u_i}$.

$$|A^{-1}b - \tilde{x}| \leq \beta v + \min((D^{-1} + vw^T)(c - \beta u), (D^{-1} + vw^T)((c - \beta u) - \Delta(I - \Delta)(c - \beta u)))$$

Proof

From the definition of Δ, it follows

$$I - \langle A \rangle D^{-1} + \Delta = uw^T. \tag{11}$$

Multiplying (11) from the left by $\langle A \rangle^{-1}$ yields

$$\langle A \rangle^{-1}(I + \Delta) = D^{-1} + vw^T. \tag{12}$$

Multiplying (12) from the right by $I - \Delta + \Delta^2$ yields

$$\langle A \rangle^{-1}(I + \Delta^3) = (D^{-1} + vw^T)(I - \Delta + \Delta^2). \tag{13}$$

Proof

From $\Delta \geq O$, we have

$$\langle A \rangle^{-1} \leq \langle A \rangle^{-1}(I + \Delta^3) \tag{14}$$

and

$$\langle A \rangle^{-1} \leq (D^{-1} + vw^T)(I - \Delta + \Delta^2). \tag{15}$$

Using (15),

$$\begin{aligned} |A^{-1}b - \tilde{x}| &\leq |A^{-1}||b - A\tilde{x}| \\ &\leq \langle A \rangle^{-1}|b - A\tilde{x}| \\ &\leq (D^{-1} + vw^T)(I - \Delta + \Delta^2)|b - A\tilde{x}| \\ &= (D^{-1} + vw^T)(c - \Delta(I - \Delta)c). \end{aligned}$$

Proof

From the definition of β and the property of H-matrix,

$$\begin{aligned} |A^{-1}b - \tilde{x}| &\leq |A^{-1}||b - A\tilde{x}| \\ &= \beta v + \langle A \rangle^{-1}(|b - A\tilde{x}| - \beta u). \end{aligned}$$

Using (4) and (15),

$$|A^{-1}b - \tilde{x}| \leq \beta v + (D^{-1} + vw^T)(c - \beta u).$$

and

$$|A^{-1}b - \tilde{x}| \leq \beta v + (D^{-1} + vw^T)((c - \beta u) - \Delta(I - \Delta)(c - \beta u)).$$

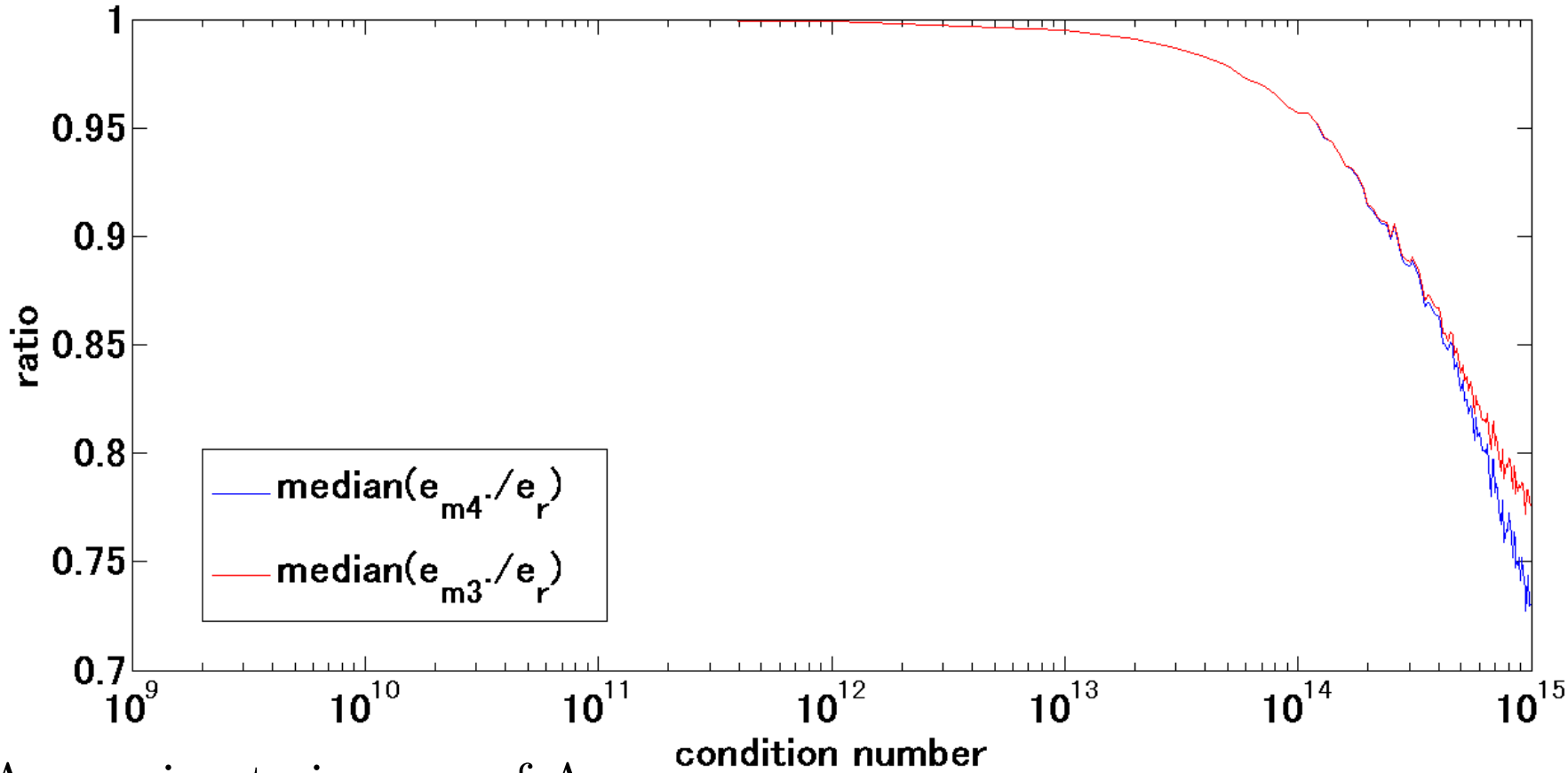
Thus,
$|A^{-1}b - \tilde{x}| \leq$
$\beta v + \min((D^{-1} + vw^T)(c - \beta u), (D^{-1} + vw^T)((c - \beta u) - \Delta(I - \Delta)(c - \beta u))).$

Numerical Result

- Test Matrix : randmat function (INTLAB function)
- Matrix Size : 100
- ϵ_r : Error bounds based on Theorem 1
  - $|A^{-1}b - \tilde{x}| \leq (D^{-1} + vw^T)|b - A\tilde{x}|$
- ϵ_{m3} : Error bounds based on (9) in Theorem 3
  - $|A^{-1}b - \tilde{x}| \leq (D^{-1} + vw^T)(c - \Delta(I - \Delta)c)$
- ϵ_{m4} : Error bounds based on (10) in Theorem 3
  - $|A^{-1}b - \tilde{x}| \leq \beta v + (D^{-1} + vw^T)((c - \beta u) - \Delta(I - \Delta)(c - \beta u))$

Numerical Result 3

- The inclusion of RA : $[\mathrm{fl}_{\triangledown}(RA), \mathrm{fl}_{\triangle}(RA)]$
- ratio := median$(\epsilon_M ./ \epsilon_R)$



R : Approximate inverse of A

$\mathrm{fl}_{\triangledown}(RA)$: setround$(-1); R * A$, $\mathrm{fl}_{\triangle}(RA)$: setround$(1); R * A$

Conclusion

- We proposed two simple modified methods
- We improved accuracy if a given matrix is an ill-conditioned matrix.
- Proposed methods is implemented in INTLAB 8.